Published online 10 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D187–D190***
doi:10.1093/nar/gkr1043

Database for bacterial group II introns

Manuel A. Candales, Adrian Duong, Keyar S. Hood, Tony Li, Ryan A. E. Neufeld, Runda Sun, Bonnie A. McNeil, Li Wu, Ashley M. Jarding and Steven Zimmerly*

Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N 1N4, Canada

Received August 30, 2011; Revised October 21, 2011; Accepted October 24, 2011

ABSTRACT

The Database for Bacterial Group II Introns (http://webapps2.ucalgary.ca/~groupii/index.html#) provides a catalogue of full-length, non-redundant group II introns present in bacterial DNA sequences in GenBank. The website is divided into three sections. The first section provides general information on group II intron properties, structures and classification. The second and main section lists information for individual introns, including insertion sites, DNA sequences, intron-encoded protein sequences and RNA secondary structure models. The final section provides tools for identification and analysis of intron sequences. These include a step-by-step guide to identify introns in genomic sequences, a local BLAST tool to identify closest intron relatives to a query sequence, and a boundary-finding tool that predicts 5′ and 3′ intron–exon junctions in an input DNA sequence. Finally, selected intron data can be downloaded in FASTA format. It is hoped that this database will be a useful resource not only to group II intron and RNA researchers, but also to microbiologists who encounter these unexpected introns in genomic sequences.

INTRODUCTION

Group II introns are a class of mobile DNAs consisting of a catalytic RNA (ribozyme) and an intron-encoded protein (IEP). The ribozyme component catalyzes self-splicing *in vitro*, at least for some introns, while the IEP promotes splicing reaction either *in vivo*, or under physiological conditions *in vitro*. The IEP also allows the intron to be mobile and insert into new genomic locations. The biochemical mechanisms for splicing and mobility reactions have been covered in detail in a number of review articles (1–9).

Despite having variable primary sequences, group II RNAs fold into a conserved secondary structure that consists of six domains (DI –DVI) emanating from a central wheel (Figure 1A) (10). Domain I is the largest domain, while domain V contains catalytic residues and domain VI contains a bulged adenosine motif that is analogous to the branchpoint of spliceosomal introns. Within a large loop of domain IV is an open reading frame (ORF) encoding the IEP (Figure 1A,B). The IEP is a multifunctional protein containing a reverse transcriptase (RT) domain that is comprised of seven sequence blocks conserved across RT families. Downstream domains include the X/thumb domain that contributes maturase (splicing) function, a DNA-binding domain (D), and an endonuclease (En) domain, with the latter two being involved in the mobility reaction. The D domain is not highly conserved in sequence among group II introns, while the En domain is absent from many introns.

Group II introns can be classified according to either ribozyme secondary structures or IEP sequences. Ribozyme structures are divided into IIA, IIB and IIC classes, based on characteristic secondary structure features and mechanisms of exon recognition (8–10). In contrast, the IEP classifications are phylogenetically based, and are denoted bacterial classes A, B, C, D, E and F, ML (mitochondrial-like), and CL1 and CL2 (chloroplast-like 1 and 2) (11–13). The two classification systems do in fact correspond, in that ML introns have IIA structures, bacterial C introns have IIC structures, and the rest have class-specific variations of IIB structures. All of these classes are present in bacteria.

Group II introns are distributed throughout eubacteria and archaebacteria, as well as chloroplasts and mitochondria of plants, fungi, protists and a few animals (2,4,14,15). Approximately a quarter of eubacterial genomes harbor at least one group II intron, whereas

*To whom correspondence should be addressed. Tel: +1 403 220 7933; Fax: +1 403 289 9311; Email: zimmerly@ucalgary.ca

The authors wish it to be known that, in their opinion, the first six authors should be regarded as joint First authors.

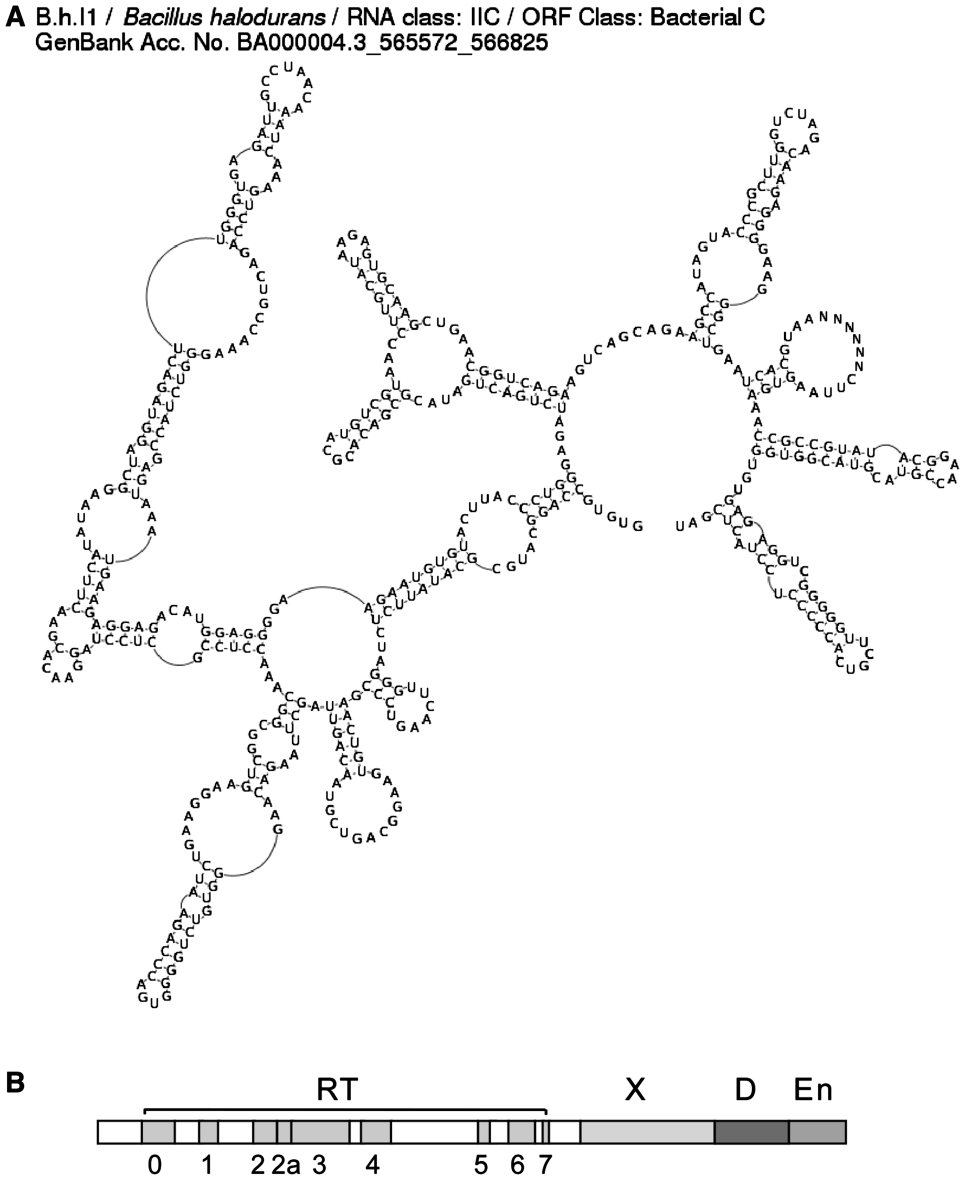

Figure 1. Group II intron structure. (**A**) An example intron RNA secondary structure, as displayed by the website. Ns in the domain IV loop denote the long ORF sequence that is not shown. (**B**) Typical ORF structure of a group II intron IEP with domains for RT, X/thumb (maturase, or splicing function), D (DNA-binding) and En (endonuclease) (drawn to scale for the Ll.LtrB intron of *Lactococcus lactis*). Some lineages of introns lack the En domain, while domain D is a functional domain and not highly conserved in sequence.

relatively few are found in archaebacteria. Despite the widespread occurrence of group II introns in bacterial genomes, they continue to be overlooked and misannotated in new sequences. A major goal of the website is to facilitate identification of introns in newly sequenced genomes by providing a reference set of correctly identified introns and varietal classes.

DATA GENERATION AND DATABASE CONTENT

The website was established in 2002 as a compendium of all bacterial group II introns in GenBank, which at that time totalled ~40 introns (16). By 2011, the number has increased to almost 400 introns in the database (http://webapps2.ucalgary.ca/~groupii/index.html#). The curation process occurs through a semi-automated series of steps that first identifies an IEP, and then locates and folds the surrounding ribozyme (to be reported elsewhere). Manual proofreading and refinements are required to maintain the quality of the results. A complication in this process is the large number of truncated and inactivated introns in bacteria, which in fact, outnumber the full-length, functional introns. Reflective of this, a major change in the curation of the database is that the main table now only lists introns that are full-length and presumably functional for both splicing and mobility. Once an intron is deemed to be full-length and functional, a name is assigned based on a species abbreviation and

intron number. Sequences that are >95% identical and in the same species are given the same name and listed only once in the table to avoid redundancy. As much as possible, names are consistent with published literature, and are not changed over time. However, name changes are inevitable when species names are changed in GenBank entries.

The database is divided into three major sections. The introductory section presents information on the basic splicing and mobility properties of group II introns, as well as RNA secondary structures, ORF structure, and the distribution and evolution of group II introns. Detailed consensus secondary structures are provided for IIA, IIB and IIC ribozyme classes, as well as the ribozymes of the IEP-based phylogenetic subclasses (ML, CL1, CL2, A, B, C, D, E and F). IEP domains are defined in a multi-sequence alignment.

In the main section, individual bacterial introns are presented in a table format, with columns denoting the intron names, species, host genes, genomic loci (e.g. plasmid, chromosome), ORF domains, ORF sizes (amino acids), ORF phylogenetic classes, RNA secondary structure classes and GenBank accession numbers. Links from accession numbers lead to corresponding GenBank entries. Links from the intron names open pages for individual introns, which show their DNA sequences, with intron and ORF boundaries denoted by colors, and also their predicted ORF sequences and RNA secondary structures. Both eubacterial and archaeal tables can be sorted by clicking on column headings. For example, one could sort the intron tables by species, ORF class, intron size, etc. to cluster introns based on those criteria.

Group II introns that do not encode ORFs are comparatively rare in bacteria genomes. Information on these introns is provided on a separate web page. Almost without exception, ORF-less introns in bacteria are found in genomes harboring a closely related ORF-containing intron, such that the IEP may act in *trans* on the ORF-less intron (13,17). It remains possible that ORF-less introns are more abundant than realized, because intron identification relies on initial identification of an IEP; however, a search for group II introns independently of the IEP did not identify significantly more ORF-less introns (13).

As previously noted, there is a large number of fragmented and/or inactivated group II introns in bacterial genomes, and we no longer attempt to document all of them. In order to represent some of these sequences, a table is shown of inactivated bacterial introns, which corresponds to an approximately complete listing as of 2005.

Group II introns in organelles differ from those in bacteria because they frequently lack IEPs and have degenerated secondary structures. The database does not display a complete listing of group II introns in mitochondria and chloroplasts. Instead, a sampling of mitochondrial and chloroplast ORF-containing introns is shown in two separate tables. Researchers are referred to GOBASE (18) and FUGOID (19) for further information on organellar introns.

TOOLS FOR IDENTIFICATION AND ANALYSIS OF GROUP II INTRONS

The third section of the database offers tools for identification and analysis of group II introns. A step-by-step guide outlines a procedure to identify introns in genomic sequences. The newly implemented BLAST search tool allows one to locate the closest intron relatives in the database to a query sequence. The search returns the top 10 hits along with alignments between the query sequence and matching intron. In analyzing a candidate intron sequence, the closest relative is quite useful for determining the correct boundaries and secondary structure. The tool can also identify truncated introns, when there is an abrupt discontinuity in the alignment.

The boundary prediction tool predicts 5′ and 3′ intron boundaries in an input DNA sequence, based on sequence profiles of the intron–exon junctions for the different subclasses of introns (20). The tool makes conservative predictions, such that boundaries identified are likely to be correct, while 5′ and 3′ boundaries may be missed for introns that do not follow the consensus strictly. A test for prediction accuracy, using known intron sequences in the database, showed correct boundary prediction for 72% of introns, incorrect boundary predictions for 3% and no prediction for 25% of introns. Regardless of the computational outcome, predicted boundaries must be confirmed by folding the intron RNAs into secondary structures, and/or by verifying that the theoretical exons, when ligated, code for a functional protein.

Finally, selected intron data can be downloaded as FASTA format text files. Downloads can be requested for intron sequence, ORF sequence (DNA or amino acid) or intron sequence with flanking sequence, and can be selected according to phylogenetic class, genus, species, intron name or accession number. For example, one could download all class ML DNA sequences, all *Bacillus* IEP amino acid sequences, or all introns present in a given GenBank entry.

Undoubtedly, the number of known group II introns will continue to grow rapidly as more genomic and metagenomic samples are sequenced. With the increased functionality and new tools provided by the intron database, it is hoped that the resource will continue to aid RNA researchers and microbiologists alike.

FUNDING

Canadian Institutes of Health Research (CIHR) (grant number MOP-93662); Natural Sciences and Engineering Research Council (NSERC) of Canada (grant number RGP 203717-02); Alberta Heritage Foundation for Medical Research (salary support, to S.Z., in part); PGS-D studentship (from NSERC, to B.A.M.). Funding for open access charge: CIHR (Canada).

Conflict of interest statement. None declared.

REFERENCES

1. Michel,F. and Feral,J. (1995) Structure and activities of group II introns. *Annu. Rev. Biochem.*, **64**, 435–461.
2. Bonen,L. and Vogel,J. (2001) The ins and outs of group II introns. *Trends Genet.*, **17**, 322–331.
3. Belfort,M., Derbyshire,V., Parker,M.M., Cousineau,B. and Lambowitz,A.M. (2002) Mobile introns: pathways and proteins. In: Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), *Mobile DNA II*. ASM Press, Washington, DC, pp. 761–783.
4. Lambowitz,A.M. and Zimmerly,S. (2004) Mobile group II introns. *Annu. Rev. Genet.*, **38**, 1–35.
5. Toro,N., Jiménez Zurdo,J.I. and García Rodríguez,F.M. (2007) Bacterial group II introns: not just splicing. *FEMS Microbiol. Rev.*, **31**, 342–358.
6. Fedorova,O. and Zingler,N. (2007) Group II introns: structure, folding and splicing mechanism. *Biol. Chem.*, **388**, 665–678.
7. Lehmann,K. and Schmidt,U. (2003) Group II introns: structural and catalytic versatility of large natural ribozymes. *Crit.Rev. Biochem. Mol. Biol.*, **38**, 249–303.
8. Michel,F., Costa,M. and Westhof,E. (2009) The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem. Sci.*, **34**, 189–199.
9. Pyle,A.M. (2010) The tertiary structure of group II introns: Implications for biological function and evolution. *Crit.Rev. Biochem. Mol. Biol.*, **45**, 215–232.
10. Michel,F., Kazuhiko,U. and Haruo,O. (1989) Comparative and functional anatomy of group II catalytic introns—a review. *Gene*, **82**, 5–30.
11. Zimmerly,S., Hausner,G. and Wu,X.-C. (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.*, **29**, 1238–1250.
12. Toro,N., Molina-Sanchez,M.D. and Fernández-López,M. (2002) Identification and characterization of bacterial class E group II introns. *Gene*, **299**, 245–250.
13. Simon,D., Clarke,N.A.C., McNeil,B.A., Johnson,I., Pantuso,D., Dai,L., Chai,D. and Zimmerly,S. (2008) Group II introns in Eubacteria and Archaea: ORF-less introns and new varieties. *RNA*, **14**, 1704–1713.
14. Vallès,Y., Halanych,K.M. and Boore,J.L. (2008) Group II introns break new boundaries: presence in a bilaterian's genome. *PLoS ONE*, **3**, e1488.
15. Copertino,D.W. and Hallick,R.B. (1993) Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem. Sci.*, **18**, 467–471.
16. Dai,L., Toor,N., Olson,R., Keeping,A. and Zimmerly,S. (2003) Database for mobile group II introns. *Nucleic Acids Res.*, **31**, 424–426.
17. Meng,Q., Wang,Y. and Liu,X.Q. (2005) An intron-encoded protein assists RNA splicing of multiple similar introns of different bacterial genes. *J. Biol. Chem.*, **280**, 35085–35088.
18. O'Brien,E.A., Zhang,Y., Wang,E., Marie,V., Badejoko,W., Lang,B.F. and Burger,G. (2009) GOBASE: an organelle genome database. *Nucleic Acids Res.*, **37**, D946–D950.
19. Li,F. and Herrin,D.L. (2002) FUGOID: functional genomics of organellar introns database. *Nucleic Acids Res.*, **30**, 385–386.
20. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.